文本分类中一种特征选择方法研究*

赵 婧，邵雄凯，刘建舟，王春枝

(湖北工业大学 计算机学院，武汉 430068)

摘 要：针对文本分类中传统特征选择方法卡方统计量和信息增益的不足进行了分析，得出文本分类中的特征选择关键在于选择出集中分布于某类文档并在该类文档中均匀分布且频繁出现的特征词。因此，综合考虑特征词的文档频、词频以及特征词的类间集中度、类内分散度，提出一种基于类内类间文档频和词频统计的特征选择评估函数，并利用该特征选择评估函数在训练集每个类别中选取一定比例的特征词组成该类别的特征词库，而训练集的特征词库则为各类别特征词库的并集。通过基于 SVM 的中文文本分类实验表明，该方法与传统的卡方统计量和信息增益相比，在一定程度上提高了文本分类的效果。

关键词：文本分类；特征选择；分散度；集中度；频度

中图分类号：TP391 doi: 10.3969/j.issn.1001-3695.2018.01.0078

Study on feature selection method in text classification

Zhao Jing, Shao Xiongkai, Liu Jianzhou, Wang Chunzhi

(*School of Computer Science Hubei University of Technology*, *Wuhan* 430068, *China*)

Abstract: The traditional feature selection method of chi-square test and information gain in text classification has its inherent defect. This paper analyzed the key of feature selection in text classification being to select feature words distributed evenly and frequently in each type of documents. This should consider not only the document frequency and term frequency of feature words, but also the inter class concentration degree and the intra class scatter degree of feature words. It proposed a feature selection evaluation function that is based on document frequency of within-class and between-class and term frequency statistics. The feature selection evaluation function could select a certain proportion of the feature words in each category of the training set to form the corresponding class of the feature word library. The entire feature word library of the training set could be composed by each of such classes as a result. It carried out the experiment of Chinese text classification based on SVM. The experimental results show that the proposed method improves the effectiveness of text classification to a certain extent, compared with the traditional chi-square test and information gain.

Key words: text classification; feature selection; distribution; concentration; frequency

0 引言

文本挖掘技术作为组织和处理海量文本数据的有效技术，近几年备受关注。文本分类作为文本挖掘中的关键技术之一，其目的是在预定义的分类体系下，根据文本的特征（内容或属性），将给定的文本与一个或多个类别相关联的过程[1]。基于机器学习的文本自动分类的整体思路大致为文本预处理；特征降维；建立文本表示模型；使用分类算法分类；分类模型评估。

特征降维作为文本分类中的重要步骤，其目的在于提高分类精度和分类效率[2]。文本通过预处理后变成由词项表示，即为原始特征空间。该原始特征空间具有高维性和稀疏性的特点，所存在的问题是：a)分类时间开销大；b)过多的特征可能会导致“维数灾难”[3]。特征降维，即将特征空间从高维降低到低维层次，从而提高分类的准确率，降低分类的时间成本。

特征降维包括特征选择（feature selection）和特征抽取（feature extraction）[3]。特征选择，即从原始特征数据集中选择出一部分具有代表性的特征。特征选择后得到的是原始特征数据集的一个子集。特征抽取，即利用原始特征空间中包含的所有信息来获得新的转换空间，从而将高维模式映射到低维模式[4]。其中，传统的特征选择方法有文档频率（document frequency，DF）、互信息（mutual information，MI）、信息增益（information gain，IG）、卡方统计量（chi-square test，CHI）

收稿日期：2018-01-31；修回日期：2018-03-20 基金项目：国家自然科学基金面上资助项目（61772180）

作者简介：赵婧（1991-），女，河南南阳人，硕士研究生，主要研究方向为文本挖掘（zjle122@163.com）；邵雄凯（1963-），男，教授，博士，主要研究方向为机器学习、数据库；刘建舟（1979-），男，讲师，硕士，主要研究方向为自然语言处理、数据库；王春枝（1963-），女，教授，博士，主要研究方向为机器学习、数据挖掘.

等[5]。Yang 等人[6]的研究结果表明，卡方统计量（CHI）和信息增益（IG）的分类效果相对较好，其结论对之后的研究具有重要的参考价值。本文主要针对 CHI 和 IG 特征选择方法进行研究和分析，并提出了一种综合考虑特征词的文档频、词频以及特征词的类间集中度、类内分散度的特征选择方法、基于类内类间文档频和词频统计（document frequency of within-class and between-class and term frequency statistics，DFCTFS）的特征选择方法。

1　相关工作

1.1　CHI 特征选择方法

CHI 以特征词 t 与类别 C_i 相互独立为前提，计算这两个变量之间的值（即偏差程度）。如果计算得到的值越大（即偏差较大），则特征词 t 与类别 C_i 越相关[7]。CHI 评估函数的公式如下：

$$\chi^2(t,c_i)=\frac{N\times(A\times D-B\times C)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{1}$$

其中： N、A、B、C、D 参数的意义如表 1 所示。

表 1　CHI 评估函数中各项参数的意义

	属于类别 Ci	不属于类别 Ci	总数
包含特征词 t 的文本数	A	B	A+B
不包含特征词 t 的文本数	C	D	C+D
总数	A+C	B+D	N=A+B+C+D

式（1）计算得到的是特征词 t 对于一个类别的 CHI 值。对于训练集为多个类别，特征词 t 对于整个训练集的 CHI 值，即为计算该特征词 t 在训练集中各类别的 CHI 值，取计算所得 CHI 值的平均值或者最大值作为结果，可用式（2）和（3）分别进行表示。

$$\chi^2_{avg}(t)=\sum_{i=1}^{M}P(C_i)\chi^2(t,C_i) \tag{2}$$

$$\chi^2_{\max}(t,C_i)=\max_{i=1}^{M}\{\chi^2(t,C_i)\} \tag{3}$$

其中：M 为类别数。

但是传统的 CHI 方法存在着不足：a）未考虑特征词在各类别中的词频分布，只考虑了特征词的文档频，导致 CHI 可能会选择文档频率高但词频低的特征词[8]，例如类别 C_i 的多数文档中都含有特征词 t，即特征词 t 在类别 C_i 的文档频率高，但特征词 t 在其每篇文档中只出现一次，即特征词 t 在类别 C_i 的词频很低，该特征词 t 并不适合代表类别 C_i，但使用 CHI 特征选择方法可能会选择该特征词 t；b）不属于该类别的特征词的干扰，因为式（1）中的因子 $(A\times D-B\times C)^2$ 的存在，导致当 BC>>AD 时，即特征词 t 不属于该类别 C_i 时，其 CHI 值也会较高，可能被选择为代表 C_i 类的特征词[9]。

1.2　IG 特征选择方法

IG 用于文本的特征选择时，衡量的是某个词的出现与否对判断一个文本是否属于该类别所提供的信息量，信息量的多少由熵来衡量。

IG 即为不考虑任何特征时文档的熵和考虑该特征后文档的熵的差值[7]。该差值表示信息不确定性的减少程度。信息不确定性减少程度越大，相应的信息增益越大；该词项提供的信息越多，该词项越重要。

因此，在进行特征选择时，通常按照 IG 值降序排列，选取一定比例的词作为特征词。IG 评估函数的公式如下：

$$IG(t)=-\sum_{i=1}^{M}P(C_i)\log P(C_i)+P(t)\sum_{i=1}^{M}P(C_i|t)\log P(C_i|t)+P(\bar{t})\sum_{i=1}^{M}P(C_i|\bar{t})\log P(C_i|\bar{t}) \tag{4}$$

其中：M 表示类别数；$P(C_i)$ 表示属于类 C_i 的文本在文本集中出现的概率，即

$$P(C_i)=\frac{\text{属于类}C_i\text{的文本数}}{\text{文本集总文本数}} \tag{5}$$

$P(t)$ 表示文本集中包含特征词 t 的文本的概率，即

$$P(t)=\frac{\text{包含词}t\text{的文本数}}{\text{文本集总文本数}} \tag{6}$$

$P(C_i|t)$ 表示文本包含特征词 t 时属于类 C_i 的条件概率，即

$$P(C_i|t)=\frac{P(C_i,t)}{P(t)}=\frac{\text{包含词}t\text{且属于类}C_i\text{的文本数}}{\text{包含词}t\text{的文本数}} \tag{7}$$

$P(C_i|\bar{t})$ 表示文本不包含特征词 t 时属于类 C_i 的条件概率，即

$$P(C_i|\bar{t})=\frac{P(C_i,\bar{t})}{P(\bar{t})}=\frac{\text{不包含词}t\text{且属于类}C_i\text{的文本数}}{\text{不包含词}t\text{的文本数}} \tag{8}$$

但是传统的 IG 方法存在着不足：a）未考虑特征词在各类别中的词频分布，只考虑了特征词的文档频，导致 IG 可能会选择文档频率高但词频低的特征词；b）考虑了特征词 t 在类别 C_i 中未出现时对于分类的贡献，但该类别未出现的特征词对特征选择也存在着干扰[10]；c）只能作全局的特征选择（指训练集中所有类别都使用相同的特征集合），而无法作本地的特征选择（指训练集中每个类别都有自己的特征集合）[11]。

2　DFCTFS 特征选择方法

2.1　DFCTFS 特征选择评估函数

综合分析 CHI 和 IG 的不足，可以得出文本分类中特征选择的关键在于选择出集中分布于某类文档并在该类文档中均匀分布且频繁出现的特征词。因此，本文综合考虑特征词的文档频、词频以及特征词的类间集中度、类内分散度，提出一种基于类内类间文档频和词频统计（DFCTFS）的特征选择方法。

2.1.1 特征词的类间集中度、类内分散度

能够代表某一类别的特征词应是集中分布在该类别中（即类间集中度高），并且在该类别中均匀分布(即类内分散度大)。综合考虑这两个因素，本文提出：

$$\alpha=\frac{\mathrm{DF}(t_k,C_i)}{\mathrm{DF}(t_k)}\bullet\frac{\mathrm{DF}(t_k,C_i)}{\mathrm{DF}(t,C_i)} \tag{9}$$

其中：$\mathrm{DF}(t_k,C_i)$ 表示特征词 t_k 在类别 C_i 中出现的文本数；$\mathrm{DF}(t_k)$ 表示特征词 t_k 在训练集所有类别中出现的文本数总和；$\mathrm{DF}(t,C_i)$ 表示类别 C_i 中所有特征词出现的文本数的总和。

其基本思想是：构造一个特征词、类别的二维矩阵。假设该二维矩阵为4*3的矩阵，如表2所示。矩阵中的元素代表特征词 t_k 在 C_i 类别出现的文本数 $DF(t_k,C_i)$。将特征词 t_k 所在行的各个 $DF(t_k,C_i)$ 相加即为 $DF(t_k)$，表示特征词 t_k 在训练集所有类别中出现的文本数总和；对特征词 t_k 所在行使用 $\frac{DF(t_k,C_i)}{DF(t_k)}$，即为计算特征词 t_k 在 C_i 类的类间集中度。将第 C_i 列的各个 $DF(t_k,C_i)$ 相加即为 $DF(t,C_i)$，表示类别 C_i 中所有特征词出现的文本数的总和；对特征词 t_k 所在列使用 $\frac{DF(t_k,C_i)}{DF(t,C_i)}$，即为计算特征词 t_k 在 C_i 类的类内分散度。

表2 特征词、类别的二维矩阵

	C_1	C_2	C_3
t_1	0	1	3
t_2	3	0	0
t_3	1	1	1
t_4	0	2	0

2.1.2 词频

能够代表某一类别的特征词应是频繁出现在该类别中（即词频较高），同时考虑对词频进行归一化处理，避免类别中的文档数对词频产生影响。因此，本文提出：

$$\beta=\frac{TF(t_k,C_i)/numDocs_i}{\sum_{i=1}^{M}[TF(t_k,C_i)/numDocs_i]} \tag{10}$$

其中：$TF(t_k,C_i)$ 表示特征词 t_k 在类别 C_i 中出现的次数；$numDocs_i$ 表示类别 C_i 的文本数；M表示类别数。

2.1.3 DFCTFS 评估函数

因此，综合考虑特征词的文档频、词频以及特征词的类间集中度、类内分散度，提出DFCTFS特征选择的评估函数，公式为

$$DFCTFS(t_k,C_i)=\alpha\bullet\beta=\frac{DF(t_k,C_i)}{DF(t_k)}\bullet\frac{DF(t_k,C_i)}{DF(t,C_i)}\bullet\frac{TF(t_k,C_i)/numDocs_i}{\sum_{i=1}^{M}[TF(t_k,C_i)/numDocs_i]} \tag{11}$$

其中：$DFCTFS(t_k,C_i)$ 表示特征词 t_k 在类别 C_i 中的DFCTFS值；$DF(t_k,C_i)$ 表示特征词 t_k 在类别 C_i 中出现的文本数；$DF(t_k)$ 表示特征词 t_k 在训练集所有类别中出现的文本数总和；$DF(t,C_i)$ 表示类别 C_i 中所有特征词出现的文本数的总和；$TF(t_k,C_i)$ 表示特征词 t_k 在类别 C_i 中出现的次数；$numDocs_i$ 表示类别 C_i 的文本数；M表示类别数。

2.2 DFCTFS 特征选择的实现思路

训练文本集通过预处理和特征选择后形成特征词库。CHI特征选择方法是依据CHI评估函数，得到每个特征词在训练集各个类别的CHI值，使用特征词在所有类别中的CHI值的平均值或者最大值作为该特征词在整个训练集中的CHI值，将所有特征词按CHI值降序排列，选取一定比例的特征词作为整个训练集的特征词库。IG特征选择方法是依据IG评估函数，得到每个特征词在整个训练集中的IG值，将所有特征词按IG值降序排列，选取一定比例的特征词作为整个训练集的特征词库。

本文提出的DFCTFS特征选择方法是依据本文提出的DFCTFS评估函数，从训练集各类别选择一定比例的特征词，将所获得的特征词取并集后形成最终的特征词库。

利用DFCTFS特征选择算法进行中文文本分类的算法步骤如图1所示。

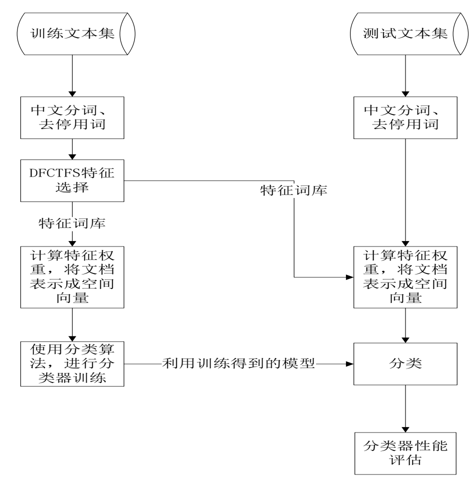


图1 中文文本分类整体流程

a）文本预处理。首先将训练集和测试集中的文本做好类别标志；其次将训练集和测试集中的文本进行分词，去停用词。文本经过预处理后变成由词项表示，即为原始特征空间。

b）特征选择。将训练集经过预处理后的原始特征空间使用DFCTFS特征选择方法得到特征词库。主要实现思路如下：

(a)构造一个特征词、类别的二维矩阵，其中行代表特征词，列代表类别，矩阵中的元素为DFCTFS值。获取训练集中原始的所有特征词，编号为 t_0～t_N。

针对训练集中的各个类别，统计特征词 t_k（k=1...N，N：词项总个数）在第 C_i（i=1...M，M：类别数）类别中出现的文本数 $DF(t_k,C_i)$ 和次数 $TF(t_k,C_i)$。根据 t_k，C_i 定位到二维矩阵相应位置，利用DFCTFS评估函数，计算 C_i 类别的特征词 t_k 的DFCTFS值，从而构造出N*M的二维矩阵。

(b)依据各类别中每个特征词的DFCTFS值，对每个类别中的特征词进行降序排列。

(c)依据文献[12]所述，对于高维的特征词空间一般选择2%～5%的特征词集合作为分类依据。根据此规则，首先获得训练集中总类别数（用M表示）以及训练集中特征词的总个数，取特征总个数的2%～5%（用numWords表示），则各类别中选择的特征词个数num为numWords除以M。

(d)各类别中都依据上一步所得num值，选取该类别中降序排列后的前num个特征词组成该类别的特征词库。

(e)得到训练集的特征词库，即为各类别所得特征词库的并集。并集，即保证特征词库中词的唯一性。

c）建立文本表示模型。其中向量空间模型使用最为广泛[13]，主要实现思路是，根据特征词库，计算训练集中每篇文本对应的特征词的权重。最常使用的权重计算方法是TF-IDF（词频-

逆文档频率），即将训练集向量化后形成一个二维矩阵，每一行代表一篇文本，每一列代表特征词库中的一个特征词。测试集作同样操作。

d）使用分类算法分类。对训练集使用分类算法进行分类器训练，得到分类模型。

e）分类器性能评估。利用训练得到的分类模型，对测试集进行分类，利用准确率、召回率、F1 值，实现对分类器的性能评价。

3　实验与分析

本文的实验在进行文本分词时使用的是中国科学院计算技术研究所研发的 ICTCLAS 汉语分词系统[14]。分词后的去停用词，使用的是哈工大停用词表[15]。

3.1　语料库

实验中使用的语料库是复旦大学计算机信息与技术系国际数据库中心自然语言处理小组整理的中文语料库[16]。选用其中的体育、历史、太空、政治、环境、经济、艺术、计算机，共 8 个类别。其中各类别文本的选取情况如表 3 所示。

表 3　语料库中训练集和测试集的选取情况

	体育	历史	太空	政治	环境	经济	艺术	计算机
训练集	400	400	400	400	400	400	400	400
测试集	65	65	65	65	65	65	65	65

3.2　分类器

实验中使用 SVM 分类算法实现中文文本分类，SVM 将基于台湾大学林智仁教授等开发的 LIBSVM 工具箱的 Java 版本。因为建立文本表示模型时使用的是向量空间模型，其本身是一个大且稀疏的矩阵，线性可分，不需要再对其进行高维映射，所以使用 SVM 中的线性核函数[17]。使用线性核函数需要寻找最优参数 C（惩罚因子）[18]。本文使用的是传统的网格搜索方法，在一定范围内，对训练集采用交叉验证的方法，找出交叉验证准确率最高的 C 值，作为 SVM 模型中的惩罚因子 C 的取值。

3.3　评价标准

实验中将采用召回率、准确率、F1 值在单个类别上进行评价。采用宏召回率、宏准确率、宏 F1 值在整体上进行评价。召回率衡量的是分类器的完备性，准确率衡量的是分类器的正确性，F1 值是调节召回率和准确率的一个平衡点。召回率 R、准确率 P、F1 值、宏召回率 MacroR、宏准确率 MacroP、宏 F1 值 MacroF1 的公式如下：

$$R=\frac{A}{A+C} \tag{12}$$

$$P=\frac{A}{A+B} \tag{13}$$

$$F1=\frac{2PR}{P+R} \tag{14}$$

$$MacroR=\frac{\sum_{i=1}^{M}R_i}{M} \tag{15}$$

$$MacroP=\frac{\sum_{i=1}^{M}P_i}{M} \tag{16}$$

$$MacroF1=\frac{\sum_{i=1}^{M}F1_i}{M} \tag{17}$$

其中：M 为类别数；A、B、C、D 参数的意义如表 4 所示。

表 4　召回率、准确率公式中的参数意义

	属于该类	不属于该类
判定为属于该类的	A	B
判定为不属于该类的	C	D

3.4　实验结果及分析

表 5～7 以及相应的图 2～4 是分别选择原始特征词集合的 2%、3%、4%，5%作为特征向量空间维数在宏召回率、宏准确率、宏 F1 值上的实验结果。通过在不同维度上对分类器性能进行整体评价，验证本文提出的 DFCTFS 特征选择方法的有效性。通过实验对比，传统的 CHI 和 IG 特征选择方法的分类宏召回率、宏准确率、宏 F1 值在不同维度上呈现出上下波动的趋势，而本文提出的 DFCTFS 特征选择方法在不同维度上的分类效果相较于 CHI 和 IG 而言，都有一定程度的提高，并且特征向量空间维数不同，提升的幅度不同。在本文的实验中，选择原始特征词集合的 5%作为特征向量空间维数，提升的幅度最大。

分析 DFCTFS 特征选择方法优于传统的 CHI 和 IG 的原因，是因为 DFCTFS 特征选择评估函数是基于类内类间文档频和词频统计的，特征选择出的特征词是集中分布于某类文档并在该类文档中均匀分布且频繁出现的。同时，DFCTFS 特征选择方法做的是本地特征选择，相较于全局特征选择方法而言，特征选择出的特征词在具体类别中更具有代表性。

表 5　CHI、IG 和 DFCTFS 在不同维度的宏召回率的比较

	2%	3%	4%	5%
CHI R	0.9288	0.925	0.9326	0.925
IG R	0.9134	0.9326	0.925	0.9307
DFCTFS R	0.9307	0.9384	0.9403	0.9461

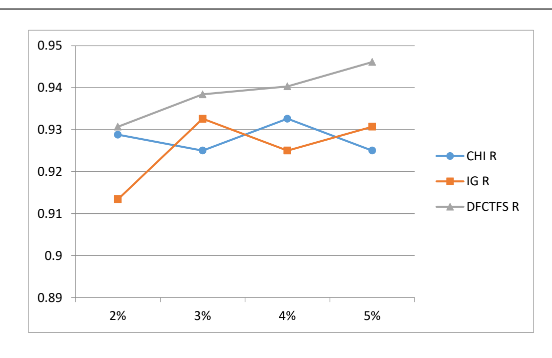


图 2　CHI、IG 和 DFCTFS 在不同维度的宏召回率的比较

表 6　CHI、IG 和 DFCTFS 在不同维度的宏准确率的比较

	2%	3%	4%	5%
CHI P	0.931	0.927	0.9343	0.9259
IG P	0.9143	0.9339	0.9271	0.9334
DFCTFS P	0.9321	0.9402	0.9416	0.947

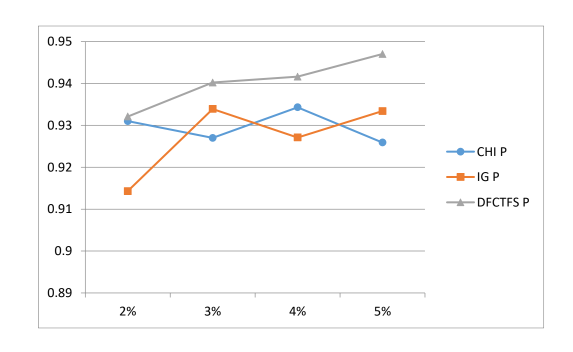


图 3 CHI、IG 和 DFCTFS 在不同维度的宏准确率的比较

表 7 CHI、IG 和 DFCTFS 在不同维度的宏 F1 值的比较

	2%	3%	4%	5%
CHI F1	0.9293	0.9252	0.9327	0.9248
IG F1	0.9132	0.9326	0.9251	0.931
DFCTFS F1	0.9309	0.9385	0.9402	0.946

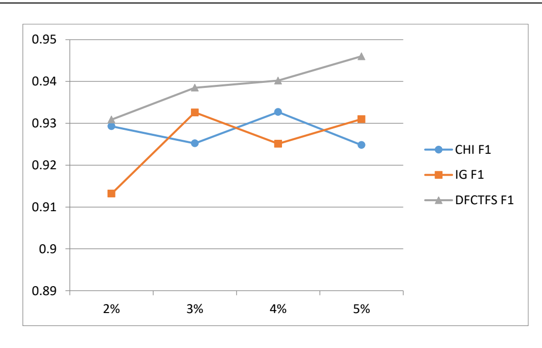


图 4 CHI、IG 和 DFCTFS 在不同维度的宏 F1 值的比较

以下实验结果是通过对具体类别分类效果的评价，进一步验证本文提出的 DFCTFS 特征选择方法的有效性。实验选择原始特征词集合的 5%作为特征向量空间维数。通过对表 8～10 及相应的图 5～7 分析，可以得出本文提出的 DFCTFS 特征选择，在所选的 8 个类别的分类效果的整体趋势上好于传统的 CHI 和 IG。通过对表 11 及相应的图 8 分析，可以得出本文提出的 DFCTFS 特征选择方法，在分类的宏召回率上与 CHI、IG 相比分别提高了 2.11%、1.54%，在宏准确率上分别提高了 2.11%、1.36%，在宏 F1 值上分别提高了 2.12%、1.5%。因此，可以得出，本文提出的 DFCTFS 特征选择方法与传统的 CHI 和 IG 相比，文本分类效果有一定程度的提高，说明了 DFCTFS 特征选择方法的有效性。

表 8 CHI、IG 和 DFCTFS 在各类别上分类召回率的比较

	CHI R	IG R	DFCTFS R
体育	0.9384	0.9538	0.9538
历史	0.8461	0.8923	0.8923
太空	0.923	0.923	0.9384
政治	0.8769	0.8615	0.923
环境	0.9846	0.9692	0.9846
经济	0.9846	0.9846	1
艺术	0.9538	0.9538	0.9538
计算机	0.8923	0.9076	0.923

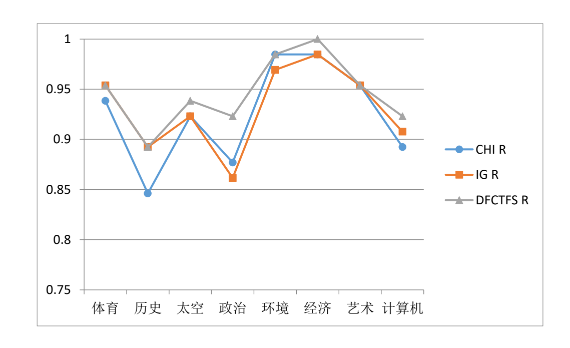


图 5 CHI、IG 和 DFCTFS 在分类召回率的比较

表 9 CHI、IG 和 DFCTFS 在各类别上分类准确率的比较

	CHI P	IG P	DFCTFS P
体育	0.9242	0.9393	0.9538
历史	0.8593	0.8529	0.9206
太空	0.8823	0.8695	0.9104
政治	0.9661	1	1
环境	0.9696	0.9692	0.9696
经济	0.9014	0.9142	0.9154
艺术	0.9841	1	0.9687
计算机	0.9206	0.9218	0.9375

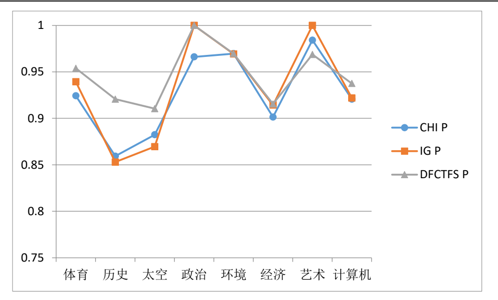


图 6 CHI、IG 和 DFCTFS 在分类准确率的比较

表 10 CHI、IG 和 DFCTFS 在各类别上分类 F1 值的比较

	CHI F1	IG F1	DFCTFS F1
体育	0.9312	0.9465	0.9538
历史	0.8527	0.8721	0.9062
太空	0.9022	0.8955	0.9242
政治	0.9193	0.9256	0.96
环境	0.977	0.9692	0.977
经济	0.9411	0.9481	0.9558
艺术	0.9687	0.9763	0.9612
计算机	0.9062	0.9147	0.9302

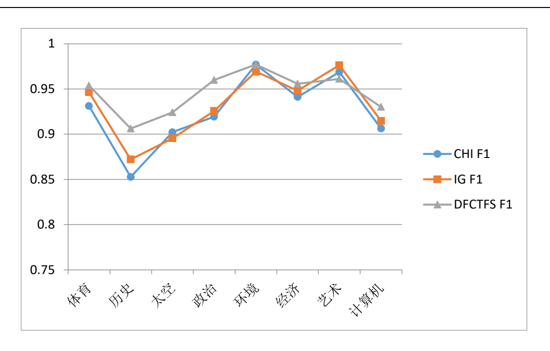


图 7 CHI、IG 和 DFCTFS 在分类 F1 值的比较

表 11 CHI、IG 和 DFCTFS 在整体分类效果上的比较

	宏 R	宏 P	宏 F1
CHI	0.925	0.9259	0.9248
IG	0.9307	0.9334	0.931
DFCTFS	0.9461	0.947	0.946

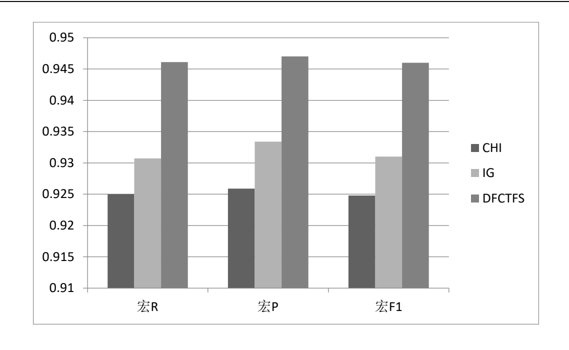


图 8 CHI、IG 和 DFCTFS 在整体分类效果上的比较

4 结束语

本文分析了传统特征选择效果较好的 CHI 和 IG 特征选择方法存在的不足，即未考虑特征词在各类别中的词频分布；类别负相关特征词的干扰；以及 IG 只能做全局的特征选择，而无法做本地的特征选择。并由此得出，文本分类中的特征选择关键在于选择出集中分布于某类文档并在该类文档中均匀分布且频繁出现的特征词。因此，综合考虑特征词的文档频、词频以及特征词的类间集中度、类内分散度，提出一种基于类内类间文档频和词频统计（DFCTFS）的特征选择方法。

通过基于 SVM 的中文文本分类实验验证，DFCTFS 特征选择与 CHI、IG 特征选择方法相比，在一定程度上提高了中文文本分类的效果。但是由于中文文本分类系统涉及文本预处理，特征降维，建立文本表示模型中的特征词权重的计算，分类算法决策等多个环节，最终的分类效果是由以上所述环节共同作用的结果。因此，仅改进特征选择的方法，只能在一定程度上提高分类的效果。

参考文献：

[1] 宗成庆. 统计自然语言处理（第 2 版）[M]. 北京：清华大学出版社，2013: 416-430. (Zong Chengqing. Statistical natural language processing (2nd ed) [M]. Beijing: Tsinghua University Press, 2013: 416-430.)

[2] 戚孝铭，施亮. 基于模拟退火及蜂群算法的优化特征选择算法 [J]. 计算机工程与设计，2013, 34 (8): 2917-2921. (Qi Xiaoming, Shi Liang. Improved feature selection algorithm based on simulated annealing algorithm and artificial bee colony algorithm [J]. Computer Engineering and Design, 2013, 34 (8): 2917-2921.)

[3] 廖一星. 文本分类及其特征降维研究 [D]. 杭州：浙江大学，2012. (Liao Yixing. Text classification and its feature reduction research [D]. Hangzhou: Zhejiang University, 2012.)

[4] Maji P, Garai P. Simultaneous feature selection and extraction using feature significance [J]. Fundamenta Informaticae, 2015, 136 (4): 405-431.

[5] Lu Yonghe, Liang Minghui, Ye Zeyuan, *et al.* Improved particle swarm optimization algorithm and its application in text feature selection [J]. Applied Soft Computing, 2015, 35: 629-636.

[6] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization [C]// Proc of the 14th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 1997: 412-420.

[7] Uysal A K, Gunal S. A novel probabilistic feature selection method for text classification [J]. Knowledge-Based Systems, 2012, 36: 226-235.

[8] 姚海英. 中文文本分类中卡方统计特征选择方法和 TF-IDF 权重计算方法的研究 [D]. 长春：吉林大学，2016. (Yao Haiying. Research on chi-square statistic feature selection method and TF-IDF feature weighting method for Chinese text classification [D]. Changchun: Jilin University, 2016.)

[9] 叶敏，汤世平，牛振东. 一种基于多特征因子改进的中文文本分类算法 [J]. 中文信息学报，2017, 31 (4): 132-137. (Ye Min, Tang Shiping, Niu Zhendong. An improved Chinese text classification algorithm based on multiple feature factors [J]. Journal of Chinese Information Processing, 2017, 31 (4): 132-137.)

[10] 董微，刘学，倪宏. 基于信息增益的自适应特征选择方法 [J]. 计算机工程与设计，2014, 35 (8): 2856-2859. (Dong Wei, Liu Xue, Ni Hong. Adaptive feature selection method based on information gain [J]. Computer Engineering and Design, 2014, 35 (8): 2856-2859.)

[11] 任永功，杨荣杰，尹明飞，等. 基于信息增益的文本特征选择方法 [J]. 计算机科学，2012, 39 (11): 127-130. (Ren Yonggong, Yang Rongjie, Yin Mingfei, *et al.* Information gain based text feature selection method [J]. Computer Science, 2012, 39 (11): 127-130.)

[12] 王小青. 中文文本分类特征选择方法研究 [D]. 重庆：西南大学，2010. (Wang Xiaoqing. Study on the selection method of Chinese text classification features [D]. Chongqing: Southwest University, 2010.)

[13] 郭正斌，张仰森，蒋玉茹. 一种面向文本分类的特征向量优化方法 [J]. 计算机应用研究，2017, 34 (8): 2299-2302. (Guo Zhengbin, Zhang Yangsen, Jiang Yuru. Feature vector optimization method for text classification [J]. Application Research of Computers, 2017, 34 (8): 2299-2302.)

[14] 中国科学院计算技术研究所. ICTCLAS2013 [EB/OL]. (2013-06-04) [2017-03-24]. http://ictclas. nlpir. org/. (Institute of Computing Technology, Chinese Academy of Sciences. ICTCLAS2013 [EB/OL]. (2013-06-04) [2017-03-24]. http://ictclas. nlpir. org/.)

[15] 哈尔滨工业大学. 哈工大停用词表扩展 [EB/OL]. (2008-05-30) [2017-04-11]. http://download. csdn. net/download/qq361277534/475580. (Harbin Institute of Technology. The expansion of the stop words in Harbin Institute of Technology [EB/OL]. (2008-05-30) [2017-04-11]. http://download. csdn. net/download/qq361277534/475580.)

[16] 复旦大学计算机信息与技术系国际数据库中心自然语言处理小组. 复旦大学中文文本分类语料库 [EB/OL]. (2011-04-21) [2017-03-24]. http://www. nlpir. org/?action-viewnews-itemid-103. (Fudan University computer information and technology department international database

center natural language processing group. Fudan University Chinese text classification corpus [EB/OL]. (2011-04-21) [2017-03-24]. http://www. nlpir. org/?action-viewnews-itemid-103.)

[17] 刘佳. 基于 SVM 的 Web 中文文本分类系统研究与实现 [D]. 西安: 西安电子科技大学, 2014. (Liu Jia. Research and implementation of categorization system for Chinese Web text based on SVM [D]. Xian: Xidian University, 2014.)

[18] Zhang Yin, Dai Miaolin, Ju Zhimin. Preliminary discussion regarding SVM kernel function selection in the twofold rock slope prediction model [J]. Journal of Computing in Civil Engineering, 2016, 30 (3): 04015031.